BLIND SEPARATION OF TEMPORALLY CORRELATED SOURCES USING A QUASI MAXIMUM LIKELIHOOD APPROACH

*Shahram HOSSEINI, Christian JUTTEN**

Laboratoire des Images et des Signaux (LIS)
46, Avenue Félix Viallet
38031 Grenoble, France.

Dinh Tuan PHAM

Laboratoire de Modélisation et Calcul
BP 53X
38041 Grenoble, France.

ABSTRACT

A quasi-maximum likelihood approach is used for separating the instantaneous mixtures of temporally correlated, independent sources without either any preliminary transformation or a priori assumption about the probability distribution of the sources. A first order Markov model is used to represent the joint probability density of successive samples of each source. The joint probability density functions are estimated from the observations using a kernel method.

1. INTRODUCTION

In this work, The Maximum Likelihood approach (ML) is used for blind separation of instantaneous mixtures of independent sources. In a general framework (without noise and with same number of sensors and sources), this problem can be formulated as follows. Having N samples of an instantaneous mixture of K sources, $\mathbf{x}(t) = \mathbf{A}\mathbf{s}(t)$, where $\mathbf{x}(t) = [x_1(t), x_2(t), \cdots, x_K(t)]^T$ and $\mathbf{s}(t) = [s_1(t), s_2(t), \cdots, s_K(t)]^T$ are respectively the vectors of the observations and of the sources and $\mathbf{A}$ is an invertible matrix, one wants to find an estimation of the matrix $\mathbf{A}$ (or its inverse, the separation matrix) up to a scale factor and a permutation.

One of the approaches which can be used consists in maximizing the likelihood function of the observations (conditioned on the matrix $\mathbf{A}$). This approach has the advantage of providing an estimator asymptotically efficient (minimum variance among non biased estimators). For the i.i.d. sources, this approach has been used by Pham and Garat [1]. They show that the separation matrix can be estimated by solving the system of equations $E[s_i\psi(s_j)] = 0$ where s_i represents the i-th source and $\psi(s_j)$ is the score function of the j-th source. In the same paper, the authors propose another method, for temporally correlated sources which consists in computing the Discrete Fourier Transform of the sources and in applying the ML approach on the results. In [2], the authors use also the ML method but they model the probability densities of the sources by a 4-th order Gram-Charlier development. Finally, in [3], the ML method is used for separating the Gaussian sources where the correlation of each source is modeled by an autoregressive model.

In this work, we study the problem in the case of temporally correlated sources and our objective is to maximize directly the likelihood function without either any preliminary transformation or a priori assumption concerning the probability density of the sources. In fact, these densities will be estimated during the maximization procedure with a kernel approach.

The remaining of the paper is organized as follows. In section 2, after the problem statement, we derive the likelihood function to be maximized, we present the method used to estimate the probability density functions from the estimation of the sources, and we propose an iterative algorithm for maximizing the derived likelihood function. In section 3, our first simulation results will be presented. Finally, in section 4, we conclude and present some perspectives.

2. METHOD

2.1. Problem statement

Having N samples of a vector $\mathbf{x}$ of dimension K, resulted than a linear transformation $\mathbf{x} = \mathbf{A}\mathbf{s}$ where $\mathbf{s}$ is a vector of independent elements and eventually correlated in the time (the sources), and $\mathbf{A}$ is a $K \times K$ invertible matrix, our objective is to find a matrix $\mathbf{B}$ so that the components of the vector $\mathbf{y} = \mathbf{B}\mathbf{x}$ are as independent as possible (independence is obtained for $\mathbf{B} = \mathbf{A}^{-1}$).

The ML method consists in maximizing the joint probability density of all the samples of all the elements of the

*This work has been partly funded by the European project BLInd Source Separation and applications (BLISS, IST 1999-14190) and by the French project Statistiques Avancées et Signal (SASI).

vector $\mathbf{x}$ (all the observations) with respect to $\mathbf{B}$:

$$f(x_1(1),\cdots,x_K(1),\cdots\cdots,x_1(N),\cdots,x_K(N)) \quad (1)$$

Considering independence of the sources, this function is equal to:

$$(\frac{1}{|det(\mathbf{B}^{-1})|})^N \prod_{i=1}^{K} f_{s_i}(\mathbf{e}_i^T\mathbf{B}\mathbf{x}(1),\mathbf{e}_i^T\mathbf{B}\mathbf{x}(2),\cdots,\mathbf{e}_i^T\mathbf{B}\mathbf{x}(N)) \quad (2)$$

where $f_{s_i}(.)$ represents the joint density of N samples of the source s_i and $\mathbf{e}_i$ is the column i of the identity matrix. Each term of $f_{s_i}(.)$ can be written as the product of the conditional densities. Even if the approach can be used in the general case, for simplifying its realization, we suppose for $f_{s_i}(.)$ a first order Markov model, *i.e.*:

$$f_{s_i}(s_i(t)|s_i(1),\cdots,s_i(t-1)) = f_{s_i}(s_i(t)|s_i(t-1)) \quad (3)$$

Equation (2) is reduced thus to:

$$(\frac{1}{|det(\mathbf{B}^{-1})|})^N \prod_{i=1}^{K}[f_{s_i}(\mathbf{e}_i^T\mathbf{B}\mathbf{x}(1)) \prod_{t=2}^{N} f_{s_i}(\mathbf{e}_i^T\mathbf{B}\mathbf{x}(t)| \mathbf{e}_i^T\mathbf{B}\mathbf{x}(t-1))] \quad (4)$$

Taking the logarithm of (4), one obtains the log-likelihood function which must be maximized to estimate the separation matrix $\mathbf{B}$. The probability densities of the sources being unknown a priori, they must be estimated from the observations. Although one can choose to estimate directly the conditional densities, we preferred to proceed by estimating the joint densities. For this purpose, after computing the logarithm of (4), using the Bayes formula, we replace the conditional densities by the joint densities divided by the marginal densities. Thus, the log-likelihood function is :

$$L = Nlog(|det(\mathbf{B})|) + \sum_{i=1}^{K}[log(f_{s_i}(\mathbf{e}_i^T\mathbf{B}\mathbf{x}(1))) + \sum_{t=2}^{N} log(\frac{f_{s_i,s_i}(\mathbf{e}_i^T\mathbf{B}\mathbf{x}(t),\mathbf{e}_i^T\mathbf{B}\mathbf{x}(t-1))}{f_{s_i}(\mathbf{e}_i^T\mathbf{B}\mathbf{x}(t-1))})] \quad (5)$$

And after the simplification, the function to be maximized becomes:

$$L = Nlog(|det(\mathbf{B})|) + \sum_{i=1}^{K}[\sum_{t=2}^{N} log(f_{s_i,s_i}(\mathbf{e}_i^T\mathbf{B}\mathbf{x}(t), \mathbf{e}_i^T\mathbf{B}\mathbf{x}(t-1))) - \sum_{t=2}^{N-1} log(f_{s_i}(\mathbf{e}_i^T\mathbf{B}\mathbf{x}(t)))] \quad (6)$$

Computing the derivative of (6) with respect to $\mathbf{B}$, it can be shown that the matrix $\mathbf{B}$ is the solution of the following system of equations[1]:

$$\sum_{t=2}^{N}[\psi_{i,i}^{(1)}(\mathbf{e}_i^T\mathbf{B}\mathbf{x}(t),\mathbf{e}_i^T\mathbf{B}\mathbf{x}(t-1))\mathbf{e}_j^T\mathbf{B}\mathbf{x}(t) + \psi_{i,i}^{(2)}(\mathbf{e}_i^T\mathbf{B}\mathbf{x}(t),\mathbf{e}_i^T\mathbf{B}\mathbf{x}(t-1))\mathbf{e}_j^T\mathbf{B}\mathbf{x}(t-1)] - \sum_{t=2}^{N-1}[\psi_i(\mathbf{e}_i^T\mathbf{B}\mathbf{x}(t))\mathbf{e}_j^T\mathbf{B}\mathbf{x}(t)] = 0 \qquad i \neq j = 1,\cdots,K \quad (7)$$

where $[\psi_{i,i}^{(1)},\psi_{i,i}^{(2)}]^T = \boldsymbol{\nabla}\log f_{s_i,s_i}(x,y)$ and $\psi_i(x) = \frac{d\log f_{s_i}(x)}{dx}$. In these last expressions, $f_{s_i,s_i}(x,y)$ represents the joint density of two successive samples of the source s_i, and $f_{s_i}(x)$ is the marginal density of the same source. The above system of equations may be solved using, for example, the Newton-Raphson adaptive algorithm. However, in this paper, we preferred to maximize directly the log-likelihood function (6) using a gradient ascent algorithm.

2.2. Estimation of the probability densities

In the above relations, the probability densities of the sources are supposed to be known, which is never the case in practice. We thus replace these densities by the densities of the separated signals $\mathbf{y} = \mathbf{B}\mathbf{x}$. Evidently, at convergence, if $\mathbf{B} = \mathbf{A}^{-1}$, these densities coincide with the densities of the sources. In order to estimate these densities, we used the kernel method. For estimating the joint densities, since the successive samples of each source are correlated, the Fukunaga formula [4] with Gaussian kernels was used. This method has the advantage of adapting to the non symmetrical data, by using only one smoothing parameter. The joint density of two successive samples of j-th component of the vector $\mathbf{y}$ is thus estimated by:

$$\hat{f}(y_j(t),y_j(t-1)) = \frac{(det\mathbf{V}_j)^{-1/2}}{(N-1)h^2}\sum_{i=2}^{N}\frac{1}{2\pi} exp(\frac{-1}{h^2}[y_j(t)-y_j(i),y_j(t-1)-y_j(i-1)]^T \mathbf{V}_j^{-1}[y_j(t)-y_j(i),y_j(t-1)-y_j(i-1)]) \quad (8)$$

In this formula, $\mathbf{V}$ represents the covariance matrix of two successive samples of y_i, and h is the smoothing parameter determining the width of each Gaussian kernel whose optimal value is: $h_{opt} = 0.96(N-1)^{-1/6}$.

For estimating the marginal densities one can proceed to integrate the joint densities. We preferred however to

[1] There are $K(K-1)$ equations. The K other equations necessary to estimate the K^2 entries of the matrix $\mathbf{B}$, are only used to remove the indeterminacy due to scale factor, and can be chosen nearly arbitrary, according to the normalization method which is used.

estimate them separately, and using following formula:

$$\hat{f}(y_j(t)) = \frac{1}{Nh}\sum_{i=1}^{N}\frac{1}{\sqrt{2\pi}}exp(\frac{-1}{2h^2}(y_j(t)-y_j(i))^2) \quad (9)$$

the optimal value of h in this last formula is: $h_{opt} = 1.06\sigma N^{-1/5}$ where σ represents the standard variation of y_j.

2.3. Algorithm

Estimation of the matrix $\mathbf{B}$ is done using a batch type iterative approach. At each step of iteration, using the current value of the matrix $\mathbf{B}$, the joint and marginal densities of the separated signals are estimated using (8) and (9) and replaced in the log-likelihood function. Afterwards, the matrix $\mathbf{B}$ is updated for maximizing this function using a gradient ascent algorithm:

$$\mathbf{B} = \mathbf{B} + \mu\nabla_{\mathbf{B}}L \quad (10)$$

The experience shows that the estimation of the densities from the samples is highly sensitive to the normalization of the data. Therefore, all the traditional methods for cancelling the indeterminacies are not equivalent. For example, if one constrain the matrix of separation $\mathbf{B}$ to have 1s on its principal diagonal, the vector $\mathbf{y}$ may take the aberrant values at the beginning of optimization, thus involving an aberrant estimate of the densities which can lead to the divergence. On the other hand, if one normalizes the rows of the matrix $\mathbf{B}$ at the beginning of each iteration, this problem does not arise any more because the vector $\mathbf{y}$ and consequently the estimates of density always take reasonable values. With this remark, and to remove the scale indeterminacy, we normalize the rows of $\mathbf{B}$ to one at the beginning of each iteration.

3. SIMULATION RESULTS

In this section, we present preliminary simulation results. Although these experiences are not sufficient for a definitive conclusion, they can confirm the relevance of the proposed method.

In the first experience, we suppose that the true sources are available and the density functions are directly estimated from the sources. In other words, we replace $y_j(i)$ by $s_j(i)$, in the relations (8) and (9) for estimating the densities.

Our experience consists in separating a mixture of two independent sources, each one representing a first order autoregressive sequence. One of the sources is Gaussian and the other uniform. The experiences were done for different sample numbers (N) and different correlation coefficients of each source (ρ_1 and ρ_2). For each chosen combination,

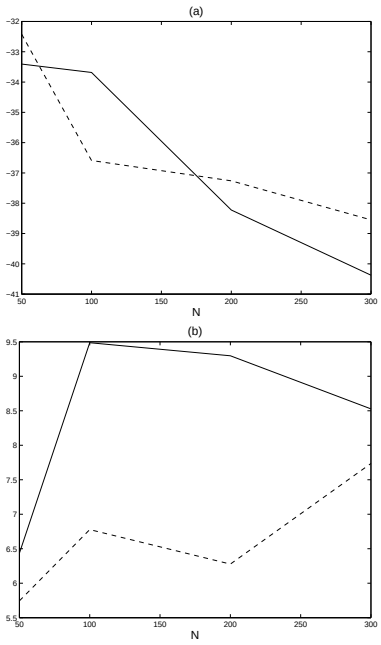

Fig. 1. (a) mean and (b) standard deviation of the residual cross-talk (in dB) for the Gaussian source (solid lines) and the uniform source (dashed lines) with respect to the number of samples (N) when the sources are supposed known .

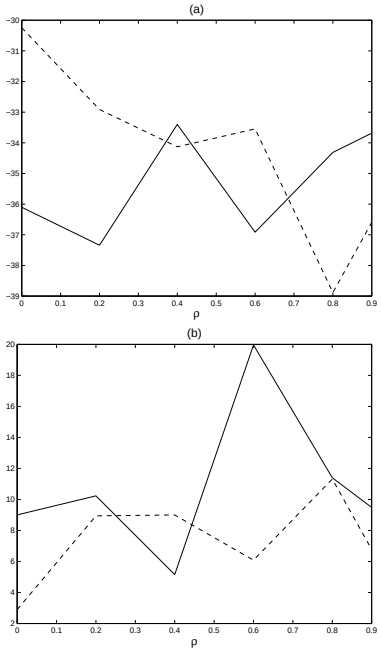

Fig. 2. (a) mean and (b) standard deviation of the residual cross-talk (in dB) for the Gaussian source (solid lines) and the uniform source (dashed lines) with respect to the correlation coefficient (ρ) when the sources are supposed known.

10 experiences, corresponding to 10 different noise seed values, are done and the mean and the standard deviation are reported. In all the experiences, the mixture matrix is:

$$\mathbf{A} = \begin{pmatrix} \sqrt{2}/2 & -\sqrt{2}/2 \\ \sqrt{2}/2 & \sqrt{2}/2 \end{pmatrix}$$

We compute the performance of the algorithm using the residual cross-talk (in dB) on the two channels:

$$C_i = 10log_{10}E[(y_i - s_i)^2] \quad i = 1, 2 \qquad (11)$$

where y_i and s_i have unit variance. The mean and the standard deviation of these criteria are shown in Figures 1 and 2 with respect to N and $\rho_1 = \rho_2$. One can remark that the performance increases with the number of samples. The cross-talks as little as -33dB for 50 samples and -40dB for 300 samples prove the good performance of the algorithm. The dependence of the performance on the correlation coefficient is not significant. It is probably due to the small number of experiences. A Monte-carlo simulation seems necessary to verify with certitude if the performance depends on the correlation coefficient or not.

In the second experiment, we repeat a similar set of simulations supposing that the sources are unknown and the densities are estimated from the observations in an iterative manner as explained in the section 2.2. The result of a sample run with $N = 100$ and $\rho_1 = \rho_2 = 0.9$ is shown in Figure 3. The mean and the standard deviation of the cross-talks for the two sources are shown in Figure 4 with respect to N. One can remark that the performance is not as good as that obtained in the first experiment but it is still acceptable for a sample size greater or equal to 200.

4. CONCLUSION AND PERSPECTIVES

In this paper, we used the maximum likelihood approach for the blind separation of the instantaneous mixtures of the temporally correlated sources without either any preliminary transformation or *a priori* assumption on the probability densities of the sources. The Gaussian kernel estimators are used to estimate the densities of the sources from the observations using an iterative algorithm. The first results confirm the relevance of the approach.

Several points could however be improved. Firstly, the algorithm converges sometimes (although rarely) toward false solutions. The experiment shows that the convergence depends mainly on the distance between the restored signals ($\mathbf{Bx}$) and the sources ($\mathbf{s}$) at the beginning of the algorithm. Convergence is almost always acquired if one replaces the estimator of densities of $\mathbf{Bx}$ by an estimator of densities of

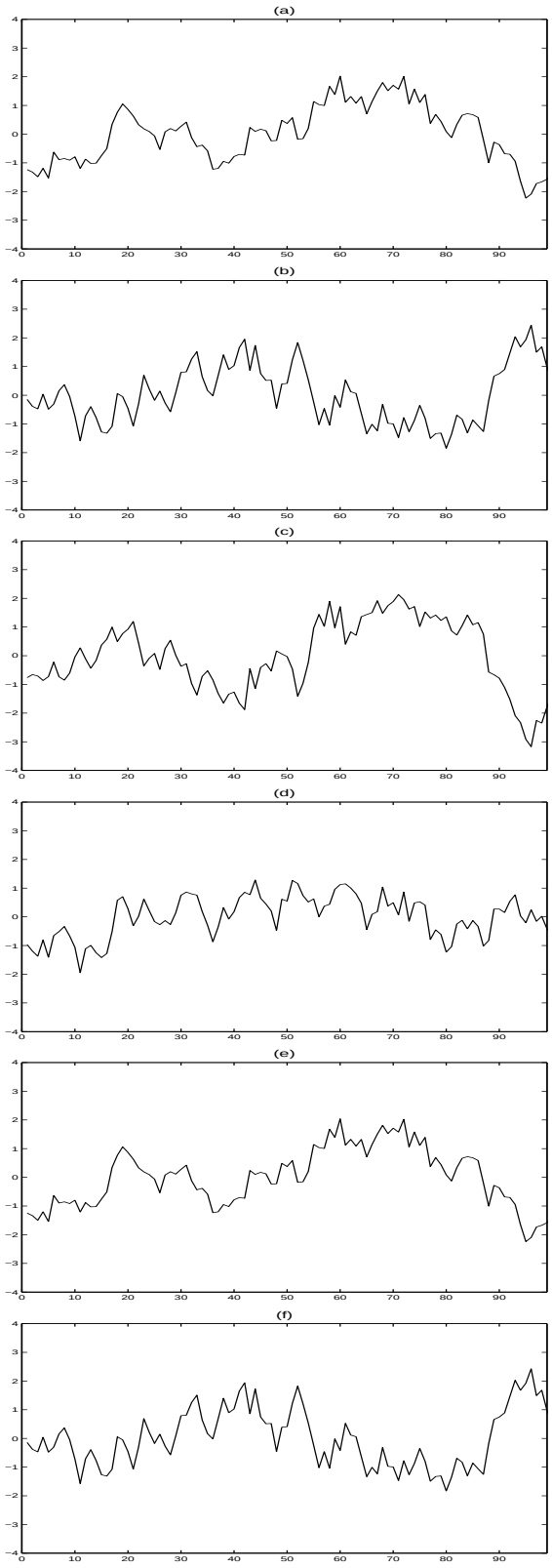

Fig. 3. (a) et (b) Sources, (c) et (d) mixtures, (e) et (f) restored signals, for $N = 100$ et $\rho_1 = \rho 2 = 0.9$ when the sources are supposed unknown.

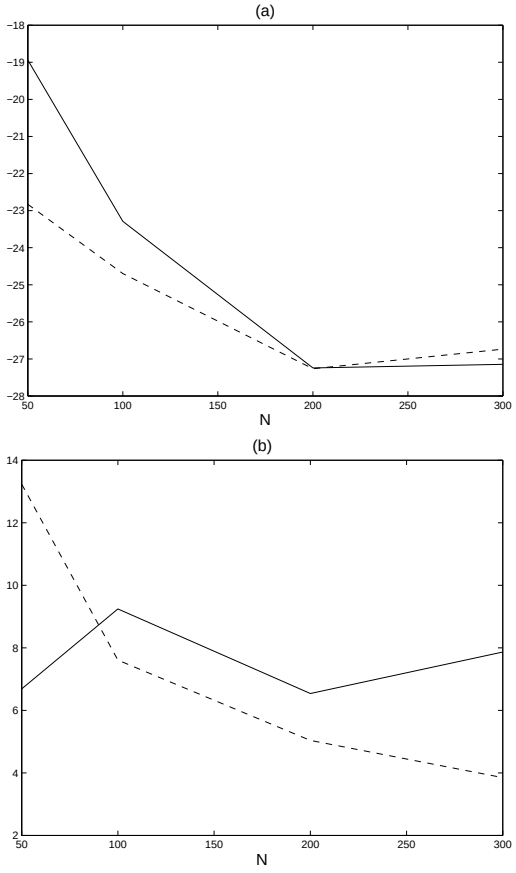

Fig. 4. (a) mean and (b) standard deviation of the residual cross-talk (in dB) for the Gaussian source (solid lines) and the uniform source (dashed lines) with respect to the number of samples (N) when the sources are supposed unknown.

sources (in other words, if one replaces $\mathbf{y}$ by $\mathbf{s}$ in the relations (8) and (9).

Secondly, the algorithm is rather slow because estimating the probability densities is time-consuming. We are currently working on less expensive, faster and more efficient algorithms for estimating these densities.

Finally, the simple gradient algorithm used in this paper is sensitive to the learning rate μ. A conjugate gradient algorithm, for example, can solve this problem.

The experiments presented in this paper had only the objective to show the interest of the method. The number of the experiences is not sufficient to derive the statistically significant results. Many other tests seem necessary to obtain a convincing conclusion. The comparison between our method and the traditional methods can permit to check if the modeling of correlation improves the performance or not. One can also test the algorithm on the correlation generated by the non linear filters. We mention that in the problem formulation, any hypothesis about the nature of temporal filters is not made, expected a first order Markov model for simplifying the realization. Finally, one can envisage to test the algorithm with the real signals, like speech or biomedical signals.

5. REFERENCES

[1] D. T. Pham et P. Garat. Blind separation of mixture of independent sources through a quasi-maximum likelihood approach. *IEEE Trans. on Signal Processing*, 45(7), July 1997.

[2] M. Gaeta et J. L. Lacoume. Estimateur du maximum de vraisemblance étendus à la séparation de sources non gaussiennes. *Traitement du Signal*, 7(5): 419-437, 1990.

[3] A. Zaïdi. *Séparation aveugle d'un mélange instantané de sources autorégressives gaussiennes par la méthode du maximum de vraisemblance*. Ph.D. Thesis, December 2000.

[4] B. W. Silverman. *Density Estimation for Statistics and Data Analysis*. Chapman and Hall, 1986.